Topologically biased random walk and community finding in networks

Vinko Zlatić,[1,2] Andrea Gabrielli,[1,3] and Guido Caldarelli[1,4,5]
[1]*Istituto Sistemi Complessi–CNR, UOS "Sapienza," Dipartimento di Fisica, Università "Sapienza," Piazzale A. Moro 2, 00185 Rome, Italy*
[2]*Theoretical Physics Division, Rudjer Bošković Institute, P.O. Box 180, HR-10002 Zagreb, Croatia*
[3]*Istituto dei Sistemi Complessi–CNR, Via dei Taurini 19, 00185 Rome, Italy*
[4]*LINKALAB, Via San Benedetto 88, 09129 Cagliari, Italy*
[5]*London Institute for Mathematical Sciences, 22 South Audley Street, Mayfair, London W1K 2NY, United Kingdom*
(Received 17 March 2010; revised manuscript received 26 October 2010; published 8 December 2010)

We present an approach of topology biased random walks for undirected networks. We focus on a one-parameter family of biases, and by using a formal analogy with perturbation theory in quantum mechanics we investigate the features of biased random walks. This analogy is extended through the use of parametric equations of motion to study the features of random walks vs parameter values. Furthermore, we show an analysis of the spectral gap maximum associated with the value of the second eigenvalue of the transition matrix related to the relaxation rate to the stationary state. Applications of these studies allow *ad hoc* algorithms for the exploration of complex networks and their communities.

DOI: 10.1103/PhysRevE.82.066109 PACS number(s): 89.75.Hc, 05.40.Fb, 02.50.Ga, 02.50.Tt

I. INTRODUCTION

The study of complex networks has notably increased in the last years with applications to a variety of fields ranging from computer science [1] and biology to social science [2–4] and finance [5]. A central problem in network science [6,7] is the study of random walks (RWs) on a graph, and in particular of the relation between the topological properties of the network and the properties of diffusion on it. This subject is not only interesting from a purely theoretical perspective, but it has also important implications to various scientific issues ranging from epidemics [8] to the classification of web pages through PAGERANK algorithm [9]. Finally, RW theory is also used in algorithms for community detection [10–14].

In this paper we set up a framework for the study of topologically biased random walks (TBRWs) on graphs. This allows us to address problems of community detection and synchronization [15] in the field of complex networks [16,17]. In particular by using topological properties of the network to bias the RWs we explore the network structure more efficiently. A similar approach but with different focus can be found in [18]. In this research we are motivated by the idea that biased random walks can be efficiently used for community finding. To this aim we introduce a set of mathematical tools which allow us an efficient investigation of the "bias parameters" space. We apply these tools to uncover some details in the spectra of graph transition matrix and use the relation between spectra and communities in order to introduce a methodology for an efficient community finding. The paper is organized as follows: in the second section we define the TBRWs. We then develop the mathematical formalism used in this paper, specifically the perturbation methods and the parametric equations of motion (PEM), to track the behavior of different biases. In the third section we focus on the behavior of spectral gap in biased random walks. We define the conditions for which such a spectral gap is maximal, and we present numerical evidence that this maximum is global. In the fourth section we present an invariant quantity for the biased random walk; such a constant quantity depends only on topology for a broad class of biased random walks. Finally, in the fifth section we present a general methodology for the application of different TBRWs in the community finding problems. We then conclude by providing a short discussion of the material presented and by providing an outlook on different possible applications of TBRWs.

II. BIASED RANDOM WALKS

RWs on graphs are a subclass of Markovian chains [19]. The traditional approach deals with the connection of the *unbiased* RW properties to the spectral features of *transition operators* associated with the network [20]. A generic graph can be represented by means of the adjacency matrix $\hat{\boldsymbol{A}}$ whose entries A_{ij} are 1 if an edge connects vertices i and j and zero, otherwise. Here, we consider undirected graphs, so that $\hat{\boldsymbol{A}}$ is symmetric. The *normal matrix* $\hat{\boldsymbol{T}}$ is related to $\hat{\boldsymbol{A}}$ through $\hat{\boldsymbol{T}}=\hat{\boldsymbol{A}}\hat{\boldsymbol{k}}^{-1}$, where $\hat{\boldsymbol{k}}$ is a diagonal matrix with $(\hat{\boldsymbol{k}})_{ii}=k_i$, i.e., the degree, or number of edges, of vertex i. In the following we use uppercase letters for nondiagonal matrices and lowercase letters for the diagonal ones. Note that by definition $k_j=\Sigma_i A_{ij}$. Consequently, $\Sigma_i T_{ij}=1$ with $T_{ij}\neq 0$ if and only if $A_{ij}=1$, i.e., if i and j are nearest-neighbor vertices. The matrix $\{T_{ij}\}$ defines the transition probabilities for an unbiased random walker to pass from j to i. In such a case T_{ij} has the same positive value for any of the neighbors i of j and vanishes for all the other vertices [21]. In analogy to the operator defining the single-step transition probabilities in the general Markovian chains, $\hat{\boldsymbol{T}}$ is also called the transition *matrix* of the unbiased RWs.

A *biased* RW on a graph can be defined by a more general transition matrix $\hat{\mathbf{T}}$, where the element T_{ij} gives again the probability that a walker on the vertex j of the graph will move to the vertex i in a single step, but depending on appropriate weights for each pair of vertex (i,j). A genuine

way to write these probabilities is to assign weights W_{ij} which represent the rates of jumps from vertex j to vertex i and normalize them:

$$T_{ij} = \frac{W_{ij}}{\sum_l W_{lj}}. \tag{1}$$

In this paper we consider biases which are self-consistently related to graph topological properties. For instance, W_{ij} can be a function of the vertex properties (the network degree, clustering, etc.), some function of the edge ones (multiplicity or shortest path betweenness), or any combination of the two. There are other choices of biases found in the literature such as, for instance, maximal entropy related biases [22]. Some of the results mentioned in this paper hold also for biases which are not connected to graph properties as will be mentioned in any such case. Our focus on graph properties for biases is directly connected with application of biased random walks in examination of community structure in complex networks.

Let us start by considering a vertex property x_i of the vertex i (it can be either local as, for example, the degree, or related to the first neighbors of i as the clustering coefficient, or global as the vertex betweenness). We choose the following form for the weights:

$$W_{ij} = A_{ij} e^{\beta x_i}, \tag{2}$$

where the parameter $\beta \in \mathbb{R}$ tunes the strength of the bias. For $\beta=0$ the unbiased case is recovered. By varying β the probability of a walker to move from vertex j to vertex i will be enhanced or reduced with respect to the unbiased case according to the property x_i of the vertex i. For instance, when $x_i=k_i$, i.e., the degree of the vertex i, for positive values of the parameter β the walker will spend more time on vertices with high degree, i.e., it will be attracted by hubs. For $\beta<0$ it will instead try to "avoid" traffic congestion by spending its time on the vertices with small degree. The entries of the transition matrix can now be written as

$$T_{ij}(\boldsymbol{x},\beta) = \frac{A_{ij} e^{\beta x_i}}{\sum_l A_{lj} e^{\beta x_l}} \equiv \frac{A_{ij} e^{\beta x_i}}{z_j(\beta)}. \tag{3}$$

For this choice of bias we find the following results: (i) we have a unique representation of any given network via operator $\hat{\boldsymbol{T}}(\boldsymbol{x},\beta)$, i.e., knowing the operator, we can reconstruct the graph; (ii) for small $|\beta|$ we can use perturbation methods around the unbiased case; (iii) this choice of bias permits us in general also to visit vertices with vanishing feature x, which instead is forbidden, for instance, for a power law $W \sim x^{\alpha}$; and (iv) this choice of biases is very common in the studies of energy landscapes, when biases represent energies $x_i \equiv E_i$ (see, for example, [23] and references therein).

In a similar way one can consider a *symmetric* edge property y_{ij} (for instance, edge multiplicity or shortest path betweenness) as bias. In this case we can write the transition probability as

$$T_{ij}(\hat{\boldsymbol{Y}},\beta) = \frac{A_{ij} e^{\beta y_{ij}}}{\sum_l A_{lj} e^{\beta y_{lj}}}. \tag{4}$$

The general case of some complicated multiparameter bias strategy can be finally written as

$$T_{ij}(\boldsymbol{x},\hat{\boldsymbol{Y}},\beta) = \frac{A_{ij} \exp\left(\sum_\nu \beta_\nu x_i^{(\nu)} + \sum_\mu \beta_\mu y_{ij}^{(\mu)}\right)}{\sum_l A_{lj} \exp\left(\sum_\nu \beta_\nu x_l^{(\nu)} + \sum_\mu \beta_\mu y_{lj}^{(\mu)}\right)}. \tag{5}$$

While we mostly consider biased RWs based on vertex properties, as shown below, most of the results can be extended to the other cases. The transition matrix in the former case can also be written as $\hat{\boldsymbol{T}}(x,\beta)=\hat{\boldsymbol{w}}\hat{\boldsymbol{A}}\hat{\boldsymbol{z}}^{-1}$, where the diagonal matrices $\hat{\boldsymbol{w}}$ and $\hat{\boldsymbol{z}}$ are such that $w_{ii}=e^{\beta x_i}$ and $z_{ii}^{-1}=1/\Sigma_l A_{li} e^{\beta x_l}$. The Frobenius-Perron theorem implies that the largest eigenvalue of $\hat{\boldsymbol{T}}(x,\beta)$ is always $\lambda_1(\beta)=1$ [19]. Furthermore, the eigenvector $\boldsymbol{v}_1$ associated with λ_1 is strictly positive in a connected aperiodic graph. Its normalized version, denoted as $\boldsymbol{p}(\beta)$, gives the asymptotic stationary distribution of the biased RWs on the graph. Assuming for it the form $p_i(\beta)=\Omega(\beta)^{-1} g_i(\beta) z_i(\beta)$, where $z_i=\Sigma_j A_{ij} e^{\beta x_j}$ and $\Omega(\beta)>0$ is a normalization constant, and plugging this in the equation $\boldsymbol{p}=\hat{\boldsymbol{T}}\boldsymbol{p}$, we get

$$p_i = \sum_j T_{ij}(\boldsymbol{x},\beta) p_j = \Omega^{-1} e^{\beta x_i} \sum_j A_{ij} g_j. \tag{6}$$

Hence, the equation holds if and only if $g_i=e^{\beta x_i}$. Therefore, the stable asymptotic distribution of vertex centered biased RWs is

$$p_i(\beta) = \Omega(\beta)^{-1} e^{\beta x_i} z_i(\beta). \tag{7}$$

For $\beta=0$ we have the usual form of the stationary distribution in an unbiased RW where $z_i(0)=k_i$ and $\Omega(0)=\Sigma_i k_i$. For general β it can be easily demonstrated that the asymptotic solution of an edge biased RW is $p_i=\Omega^{-1} z_i$, while for a multiparametric RW the solution is $p_i=\Omega^{-1}[\exp(\Sigma_\nu \beta_\nu x_i^{(\nu)})] z_i$.

Using Eqs. (7) and (3) we can prove that the detailed balance condition $T_{ij}p_j=T_{ji}p_i$ holds. At this point it is convenient to introduce a different approach to the problem [10]. We start by symmetrizing the matrix $\hat{\boldsymbol{T}}(\boldsymbol{x},\beta)$ in the following way:

$$\hat{\boldsymbol{T}}^s(\boldsymbol{x},\beta) = [\hat{\boldsymbol{p}}(\beta)]^{-1/2} \hat{\boldsymbol{T}}(\boldsymbol{x},\beta) [\hat{\boldsymbol{p}}(\beta)]^{1/2}, \tag{8}$$

where $\hat{\mathbf{p}}(\beta)$ is the diagonal matrix with the stationary distribution $\{p_i(\beta)\}$ on the diagonal. The entries of the symmetric matrix for the vertex centered case are given by

$$T_{ij}^s(\boldsymbol{x},\beta) = T_{ji}^s(\boldsymbol{x},\beta) = A_{ij} \frac{e^{(1/2)\beta(x_i+x_j)}}{\sqrt{z_i z_j}}. \tag{9}$$

The symmetric matrix $\hat{\boldsymbol{T}}^s(\boldsymbol{x},\beta)$ shares the same eigenvalues with the matrix $\hat{\boldsymbol{T}}(\boldsymbol{x},\beta)$; anyhow the set of eigenvectors is different and forms a complete orthogonal basis, allowing us to define a meaningful "distance" between vertices. Such a distance can provide important additional information in the

problem of community partition of complex networks. If $\mathbf{v}_\nu$ is the νth eigenvector of the asymmetric matrix $\hat{\boldsymbol{T}}(\boldsymbol{x},\beta)$ associated with the eigenvalue $\lambda_\nu(\beta)$ (therefore, $\mathbf{v}_1=\mathbf{p}$), the corresponding eigenvector $|v_\nu\rangle$ of the symmetric matrix $\hat{\boldsymbol{T}}^s(\boldsymbol{x},\beta)$ can always be written as $|v_\nu\rangle_i=v_{\nu,i}/\sqrt{p_i}$. In particular for $\nu=1$ we have $|v_1\rangle_i\equiv|p\rangle_i=\sqrt{p_i}$. The same transformation (8) can be applied to the most general multiparametric RWs. In that case the symmetric operator is

$$T^s_{ij}(\boldsymbol{x},\beta)=T^s_{ji}(\boldsymbol{x},\beta)=A_{ij}\frac{\exp\left[\sum_\nu \frac{\beta_\nu}{2}(x_i^{(\nu)}+x_j^{(\nu)})+\sum_\mu \beta_\mu y_{ij}^{(\mu)}\right]}{\sqrt{z_i z_j}}. \tag{10}$$

This form also enables the usage of perturbation theory for Hermitian linear operators. For instance, knowing the eigenvalue $\lambda_\nu(\beta)$ associated with eigenvector $|v_\nu(\beta)\rangle$, we can write the following expansions at sufficiently small $\Delta\beta$: $\lambda_\nu(\beta+\Delta\beta)=\lambda_\nu(\beta)+\Delta\beta\lambda_\nu^{(1)}(\beta)+\cdots$ and $|v_\nu(\beta+\Delta\beta)\rangle=|v_\nu^{(0)}(\beta)\rangle+\Delta\beta|v_\nu^{(1)}(\beta)\rangle+\cdots$. It follows that for a vertex centered bias

$$\lambda_\nu^{(1)}(\beta)=\langle v_\nu(\beta)|\hat{\boldsymbol{T}}^{s(1)}(\boldsymbol{x},\beta)|v_\nu(\beta)\rangle, \tag{11}$$

where

$$\hat{\boldsymbol{T}}^{s(1)}(\boldsymbol{x},\beta)\equiv\frac{\partial\hat{\boldsymbol{T}}^s(\boldsymbol{x},\beta)}{\partial\beta}=\frac{1}{2}[\{\hat{\boldsymbol{x}},\hat{\boldsymbol{T}}^s\}_+-\{\hat{\tilde{\boldsymbol{x}}}(\beta),\hat{\boldsymbol{T}}^s\}_+], \tag{12}$$

with $\{\cdot,\cdot\}_+$ being the anticommutator operator. The operator $\hat{\boldsymbol{x}}$ and $\hat{\tilde{\boldsymbol{x}}}(\beta)$ are diagonal matrices with $(\hat{\boldsymbol{x}})_{ii}=x_i$ and $\hat{\tilde{\boldsymbol{x}}}_{ii}(\beta)=\Sigma_l A_{li}x_l e^{\beta x_l}/z(i)$, which is the expected value of x that a random walker will find moving from vertex i to its neighbors. In the case of edge bias the change of symmetric matrix with parameter β can be written as $\partial\hat{\boldsymbol{T}}^s(\beta)/\partial\beta=\hat{\boldsymbol{Y}}\star\hat{\boldsymbol{T}}^s(\beta)-1/2\{\hat{\tilde{\boldsymbol{y}}},\hat{\boldsymbol{T}}^s(\beta)\}_+$, where $\star$ represents the Schur-Hadamard product, i.e., elementwise multiplication of matrix elements. The eigenvector components in $\beta+\Delta\beta$ at the first order of expansion in the basis of the eigenvectors at β are given by (for $\mu\neq\nu$)

$$\langle v_\mu(\beta)|v_\nu^{(1)}(\beta)\rangle=\frac{\langle v_\mu(\beta)|\hat{\boldsymbol{T}}^{s(1)}(\beta)|v_\nu(\beta)\rangle}{\lambda_\mu(\beta)-\lambda_\nu(\beta)}. \tag{13}$$

For $\mu=\nu$ the product $\langle v_\mu(\beta)|v_\nu^{(1)}(\beta)\rangle$ vanishes and Eqs. (12) and (13) hold only for nondegenerate cases. In general, usual quantum-mechanical perturbation theory can be used to go to higher-order perturbations or to take into account degeneracy of eigenvalues.

We can also exploit further the formal analogy with quantum mechanics using PEM [24,25] to study the β dependence of the spectrum of $\hat{\boldsymbol{T}}^s$. If we know such spectrum for one value of β, we can calculate it for any other value of β by solving a set of differential equations corresponding to PEM in quantum mechanics. They are nothing else but the expressions of Eqs. (11) and (13) in an arbitrary complete orthonormal base $\{|\phi_\nu\rangle\}$. First the eigenvector is expanded in such a base: $|v_\nu(\beta)\rangle=\Sigma|\phi_\xi\rangle\langle\phi_\xi|v_\nu(\beta)\rangle\equiv\Sigma_\xi c_{\nu\xi}(\beta)|\phi_\xi\rangle$. We can then write

$$\frac{\partial\lambda_\nu}{\partial\beta}=\boldsymbol{c}_\nu^\top(\beta)\frac{\partial\hat{\boldsymbol{T}}^{s,\phi}(\beta)}{\partial\beta}\boldsymbol{c}_\nu(\beta), \tag{14}$$

where $\boldsymbol{c}_\nu$ ($\boldsymbol{c}_\nu^\top$) is a column (row) vector with entries $c_{\nu\xi}(\beta)$ and $\hat{\boldsymbol{T}}^{s,\phi}$ is the matrix with entries $\hat{T}^{s,\phi}_{\nu\xi}(\beta)\equiv\langle\phi_\nu|\hat{\boldsymbol{T}}^s(\beta)|\phi_\xi\rangle$. Let us now define the matrix $\hat{\boldsymbol{N}}(\beta)$ whose rows are the copies of vector $\boldsymbol{c}_\nu^\top(\beta)$. The differential equation for the eigenvectors in the basis $\{|\phi_\nu\rangle\}$ is then [25]

$$\frac{\partial\boldsymbol{c}_\nu(\beta)}{\partial\beta}=[\hat{\boldsymbol{T}}^{s,\phi}(\beta)-\lambda_\nu(\beta)+\hat{\boldsymbol{N}}(\beta)]^{-1}\times\left(\boldsymbol{c}_\nu^\top(\beta)\frac{\partial\hat{\boldsymbol{T}}^{s,\phi}(\beta)}{\partial\beta}\boldsymbol{c}_\nu(\beta)-\frac{\partial\hat{\boldsymbol{T}}^{s,\phi}(\beta)}{\partial\beta}\right)\boldsymbol{c}_\nu(\beta). \tag{15}$$

A practical way to integrate Eqs. (14) and (15) can be found in [25]. In order to calculate the parameter dependence of eigenvectors and eigenvalues, the best way to proceed is to perform an LU decomposition of the matrix $[\hat{\boldsymbol{T}}^{s,\phi}(\beta)-\lambda_\nu(\beta)+\hat{\boldsymbol{N}}(\beta)]^{-1}$ as the product of a lower triangular matrix $\hat{\boldsymbol{L}}$ and an upper triangular matrix $\hat{\boldsymbol{U}}$, and integrate differential equations of higher order which can be constructed in the same way as Eqs. (14) and (15) [25]. A suitable choice for the basis is just the ordinary unit vectors spanned by vertices, i.e., $\{|\phi\rangle\}\equiv\{|e\rangle\}$. We found that for practical purposes, depending on the studied network, it is appropriate to use PEM until the error increases too much and then diagonalize matrix again to get a better precision. PEM efficiently enables the study of large sets of parameters for large networks due to its competitive advantage over ordinary diagonalization.

III. SPECTRAL GAP

A key variable in the spectral theory of graphs is the *spectral gap* $\mu=(\lambda_1-\lambda_2)$, i.e., the difference between first unitary and the second eigenvalues. The spectral gap measures how fast the information on the RW initial distribution is destroyed and the stationary distribution is approached. The characteristic time for that is $\tau=-1/\ln[(1-\mu)]\simeq 1/\mu$ [10]. We show in Fig. 1 the dependence of spectral gap of simulated graphs with communities for different strategies (degree, clustering, and multiplicity based) at a given value of parameter β. In all investigated cases the spectral gap has its well-defined maximum, i.e., the value of parameter β for which the random walker converges to stationary distribution with the largest rate.

The condition of maximal spectral gap implies that it is a stationary point for the function $\lambda_2(\beta)$, i.e., that its first-order perturbation coefficient vanishes at this point:

$$0=\langle v_2(\beta_m)|\frac{\partial\hat{\boldsymbol{T}}^s(\beta_m)}{\partial\beta}|v_2(\beta_m)\rangle=\langle v_2(\beta_m)|[\hat{\boldsymbol{x}}-\hat{\tilde{\boldsymbol{x}}}(\beta_m)]|v_2(\beta_m)\rangle, \tag{16}$$

where $\hat{\boldsymbol{x}}$ and $\hat{\tilde{\boldsymbol{x}}}$ are defined above. The squares of entries, $c_{2,i}^2(\beta)$, of the vector $|v_2(\beta)\rangle$ in the chosen basis $|\phi_i\rangle\equiv|e\rangle$

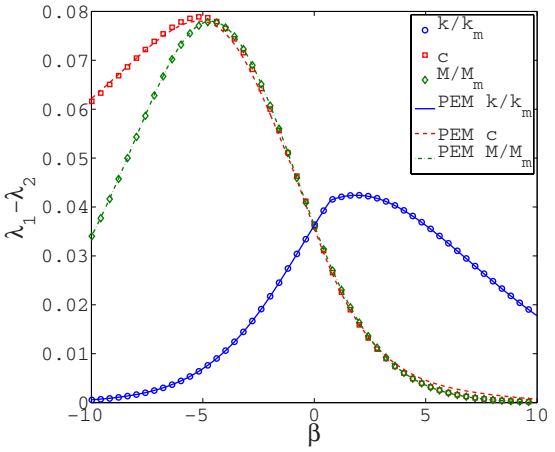

FIG. 1. (Color online) Plot of the spectral gap $\lambda_1-\lambda_2$ vs β for networks of ten communities with ten vertices each (the probability for an edge to be in a community is $p_i=0.3$, while outside the community it is $p_o=0.05$). Solid points represent the solutions computed via diagonalization, while lines report the value obtained through integration of PEM. Different bias choices have been tested. Circles (blue) are related to degree-based strategy, squares (red) are related to clustering-based strategies, and diamonds (green) are related to multiplicity-based strategies. The physical quantities to get the variable x in Eq. (2) in these strategies have been normalized with respect to their maximum values.

define a particular measure on the graph. Equation (16) can be written as $\Sigma_i c_{2,i}^2(\beta_m)(x_i-\bar{x}_i(\beta_m))=0$.

Thus, we conclude that the local spectral maximum is achieved if the average difference between property x_i and its expectation $\bar{x}_i$, with respect to this measure, in the neighborhood of vertex i vanishes. We have studied the behavior of spectral gap for different sets of real and simulated networks (Barabási-Albert model with different ranges of parameters, Erdős-Rényi model, and random networks with given community structure) and three different strategies (degree based, clustering based, and multiplicity based). Although in general it is not clear that the local maximum of spectral gap is unique, we have found only one maximum in all the studied networks. This observation is interesting because for all cases the shapes of spectral gap vs β look typically Gaussian-like. In both limits $\beta\to\pm\infty$ the spectral gap of heterogeneous network is indeed typically zero, as the RW stays in the vicinity of the vertices with maximal or minimal value of studied property x_i.

IV. RANDOM-WALK INVARIANT

A fundamental question in the theory of complex networks is how topology affects dynamics on networks. Our choice of β-parametrized biases provides a useful tool to investigate this relationship. A central issue is, for instance, given by the search of properties of the transition matrix $\boldsymbol{T}$, which are independent of β and the chosen bias, but depend only on the topology of the network. An important example comes from the analysis of the determinant of $\boldsymbol{T}$ as a function of the bias parameters:

$$\frac{\partial\prod_{\mu=1}^{N}\lambda_\mu}{\partial\beta}=\sum_{\mu=1}^{N}\langle\mu|\frac{\partial\boldsymbol{T}^s}{\partial\beta}|\mu\rangle\prod_{\nu\neq\mu}\lambda_\nu. \tag{17}$$

For vertex centered bias using Eq. (12) we have

$$\frac{\partial\prod_{\mu=1}^{N}\lambda_\mu}{\partial\beta}=\sum_{\mu=1}^{N}\langle\mu|(\hat{\boldsymbol{x}}-\hat{\bar{\boldsymbol{x}}})|\mu\rangle\prod_{\nu=1}^{N}\lambda_\nu, \tag{18}$$

and using the diagonality of $\hat{\boldsymbol{x}}$ and $\hat{\bar{\boldsymbol{x}}}_{ii}(\beta)=\partial\ln z_i/\partial\beta$,

$$\prod_{\mu=1}^{N}\frac{\lambda_\mu(\beta)z_\mu(\beta)}{e^{\beta x_\mu}}=\prod_{\mu=1}^{N}\frac{\lambda_\mu(\beta_0)z_\mu(\beta_0)}{e^{\beta_0 x_\mu}}. \tag{19}$$

In other words the quantity $\Pi_{\mu=1}^{N}\lambda_\mu(\beta)z_\mu(\beta)/e^{\beta x_\mu}$ is a topological constant which does not depend on the choice of parameters. For $\beta=0$ we get $\Pi_{\mu=1}^{N}\lambda_\mu k_\mu=$const, and it follows that this quantity does not depend on the choice of vertex biases x_i either. It can be shown that such a quantity coincides with the determinant of adjacency matrix which must be conserved for all processes.

V. COMMUNITY FINDING

There are many competing algorithms and methods for community detection [11]. Despite a significant scientific effort to find such reliable algorithms, there is no agreement yet on a single general solving algorithm for the various cases. In this section instead of adding another precise recipe, we want to suggest a general methodology based on TBRWs which could be used for community detection algorithms. To add trouble, the very definition of communities is not a solid one. In most of the cases we define communities as connected subgraphs whose density of edges is larger within the proposed community than outside it (a concept quantified by modularity [14]).

The scientific community is therefore thriving to find a benchmark in order to assess the success of various methods. One approach is to create synthetic graphs with assigned community structure (benchmark algorithms) and test through them the community detection recipes [26]. The Girvan-Newman (GN) [14] and Lancichinetti-Fortunato-Radicchi (LFR) [27] are the most common benchmark algorithms. In both these models several topological properties (not only edge density) are unevenly distributed within the same community and between different ones. We use this property to propose a methodology creating suitable TBRWs for community detection. The difference between internal and external parts of a community is related to the "physical" meaning of the graph. In many real processes the establishment of a community is facilitated by the subgraph structure. For instance, in social networks agents have a higher probability of communication when they share a lot of friends. We test our approach on GN benchmark since in this case we can easily compute the expected differences between the frequencies of biased variables within and outside the community.

In this section we will describe how to use TBRWs for community detection. For $\beta=0$ our method is rather similar to the one introduced by Donetti and Muñoz [29]. The most notable difference is that we consider the spectral properties of transition matrix instead of the Laplacian one. We decide if a vertex belongs to a community according to the following ideas: (i) we expect that the vertices belonging to the same community have similar values of eigenvector components and (ii) we expect relevant eigenvectors to have the largest eigenvalues [28]. Indeed, spectral gap is associated with temporal convergence of random walker fluctuations to the ergodic stationary state. If the network has well-defined communities, we expect the random walker to spend some time in the community rather than escaping immediately out of it. Therefore, the speed of convergence to the ergodic state should be related to the community structure. Therefore, eigenvectors associated with largest eigenvalues (except for the maximal eigenvalue 1) should be correlated with community structure. Going back to the above-mentioned Donetti and Muñoz approach here we use the fact that some vertex properties will be more common inside a community and less frequent between different communities. We then vary the bias parameters trying both to shrink the spectral gap in transition matrix and to maximize the separation between relevant eigenvalues and the rest of the spectra.

For example, in the case of GN benchmark the network consists of four communities each with $n=32$ vertices, i.e., $N=128$ vertices all together. The probability that the two vertices which belong to the same community are connected is p_{in}. The probability that the two vertices which belong to different communities are connected is p_{out}. The fundamental parameter [11] which characterizes the difficulty of detecting the structure is

$$\mu = \frac{\bar{k}_{out}}{\bar{k}_{out} + \bar{k}_{in}}, \tag{20}$$

where $\bar{k}_{out}=p_{out}(N-n)$ is the mean degree related to intercommunity connections and $\bar{k}_{in}=p_{in}(n-1)$ is the mean degree related to edges inside community. As a rule of thumb we can expect to find well-defined communities when $\mu<1/2$ and observe some signature of communities even when $\mu<3/4$ [26]. The probabilities p_{in} and p_{out} are related via the control parameter μ as $p_{out}=[(n-1)\mu/(N-n)(1-\mu)]p_{in}$.

We now examine the edge multiplicity. The latter is defined as the number of common neighbors shared by neighboring vertices. The expected multiplicities of an edge connecting vertices of intercommunity and inside communities are, respectively,

$$E(M_{out}) = 2p_{in}p_{out}(n-1) + p_{out}^2(N-2n),$$

$$E(M_{in}) = p_{in}^2(n-2) + p_{out}^2(N-n). \tag{21}$$

In Fig. 2 we plot the ratio of the quantities above defined, $E(M_{out})/E(M_{in})$, vs the parameter μ.

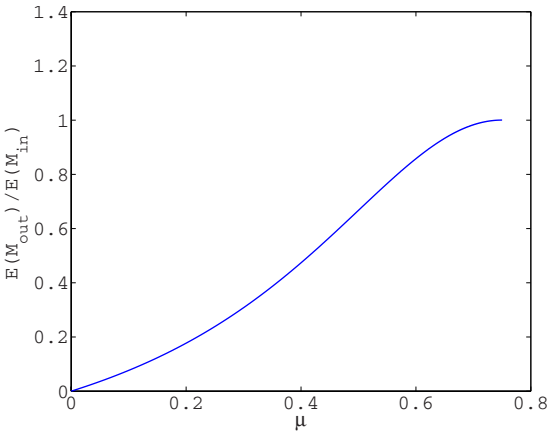

FIG. 2. (Color online) The ratio of expected value of multiplicity for edges that are connecting vertices in different communities to the expected value of multiplicity for edges that are connecting vertices in the same community with respect to parameter μ.

We see that even for $\mu>0.5$ the ratio remains smaller than 1 implying that the multiplicity is more common in the edges in the same community. Based on this analysis for this particular example we expect that if we want to find well-defined communities via TBRWs we have to increase bias with respect to the multiplicity. Through numerical simulations we find that the number of communities is related to number of eigenvalues in the "community band." Namely, one in general observes a gap between eigenvalues $\lambda_2,\ldots,\lambda_{N/n-1}$ and the next eigenvalue evident in a network with a strong community structure ($\mu \ll 1/2$). The explanation that we give for that phenomenon can be expressed by considering a network of n separated graphs. For such a network there are n degenerate eigenvalues $\lambda_1 \cdots \lambda_n=1$. If we now start to connect these graphs with very few edges, such a degeneracy is broken with the largest eigenvalue remaining 1 while the next $(n-1)$ eigenvalues staying close to it. The distance between any two of this set of $(n-1)$ eigenvalues will be smaller than the gap between this community band and the rest of the eigenvalues in the spectrum. Therefore, the number of eigenvalues different from 1 which are forming this community band is always equal to the number of communities minus 1, at least for different GN-type networks with different numbers of communities and different sizes, as long as $\mu \ll 1/2$. For example, in the case of 1000 GN networks described with parameters $N=128$, $n=32$, $p_{in}=0.35$, and $p_{out}=0.05$, i.e., $\mu=0.125$, the histograms of eigenvalues are depicted in Fig. 3.

For our purposes we used two-parameter biased RWs, in which topological properties are $x_i \equiv k_i/\max(k)$, i.e., the normalized degree (with respect to maximal degree in the network) and $y_{ij}=M_{ij}/\max(M_{ij})$, i.e., the normalized multiplicity (with respect to maximal multiplicity in the network). We chose GN network whose parameters are $N=128$, $n=32$, $p_{in}=16/62$, and $p_{out}=1/12$, for which $\mu=1/2$. With $N/n=4$ being the number of communities, as a criterion for good choice of parameters, we decided to use the difference between λ_4 and λ_5, i.e., we decided to maximize the gap between community band and the rest of eigenvalues, checking at the same time that the spectral gap shrinks. In Fig. 4, we

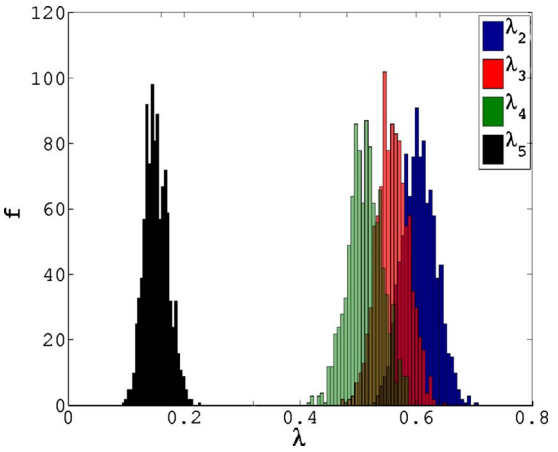

FIG. 3. (Color online) Histogram of second, third, fourth, and fifth eigenvalues of nonbiased RWs for 1000 GN networks with parameters $N=128$, $n=32$, $p_{in}=0.35$, and $p_{out}=0.05$. There is a clear gap between "community" band and the rest of the eigenvalues.

plot such a quantity with respect to different biases.

It is important to mention that for every single network instance there are different optimal parameters. This can be seen in Fig. 5, where we show the difference between unbiased and biased eigenvalues for 1000 GN nets created with same parameters. As shown in the figure the difference between fourth and fifth eigenvalues is now not necessarily the optimal for this choice of parameters. Every realization of the network should be independently analyzed, and its own parameters should be carefully chosen.

In Figs. 6 and 7 we present instead the difference between unbiased and biased projections on three eigenvectors with largest nontrivial eigenvalues. Using three-dimensional view it is easy to check that communities are better separated in the biased case than in the nonbiased case.

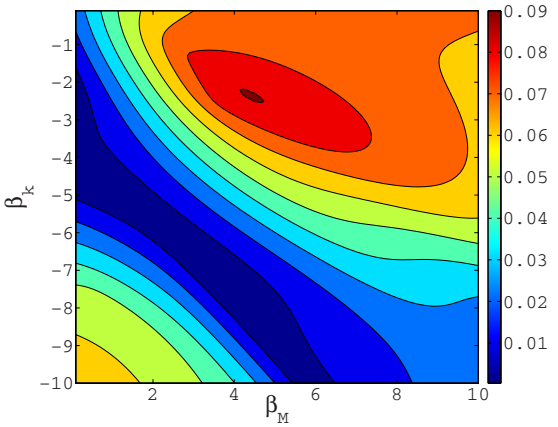

FIG. 4. (Color online) Contour plot of the difference between fourth and fifth eigenvalues $\lambda_4-\lambda_5$ as a function of parameter β_k which biases RWs according to degrees of the vertices and parameter β_M which bias RWs according to multiplicities of the edges. Both degrees and multiplicity values are normalized with respect to the maximal degree and multiplicity (therefore, the largest value is 1).

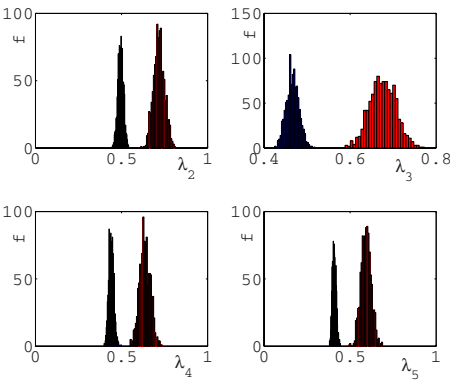

FIG. 5. (Color online) Histograms of λ_2, λ_3, λ_4, and λ_5 for 1000 GN networks described with parameters $N=128$, $n=32$, $p_{in}=16/62$, and $p_{out}=1/12$. With black color we indicate the eigenvalues of nonbiased RWs, while with red we indicate the eigenvalues of RWs biased with parameters $\beta_k=-2.5$ and $\beta_M=4.3$. Note how this choice of parameters does not maximize "community gap" for all the different realizations of monitored GN network.

VI. CONCLUSION

In this paper we presented a detailed theoretical framework to analyze the evolution of TBRWs on a graph. Using as bias some topological property of the graph itself allows us to use the RW as a tool to explore the environment. This method maps vertices of the graph to different points in the N-dimensional Euclidean space naturally associated with the given graph. In this way we can measure distances between vertices depending on the chosen bias strategy and bias parameters. In particular we developed a perturbative approach to the spectrum of eigenvalues and eigenvectors associated with the transition matrix of the system. More generally we generalized the quantum PEM approach to the present case. This led naturally to studying the behavior of the gap between the largest and the second eigenvalues of the spectrum characterizing the relaxation to the stationary Markovian

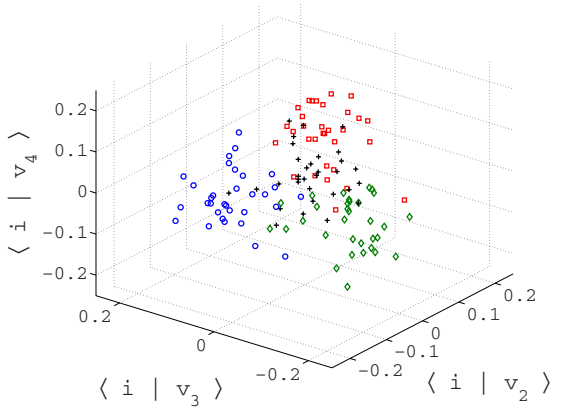

FIG. 6. (Color online) Plot of the eigenvector components of the second, third, and fourth eigenvectors. Different markers represent four different predefined communities. This is an example of GN graph with $p_{in}=16/62$ and $p_{out}=1/12$. For this choice of parameters $\mu=1/2$. There is a strong dispersion between different vertices which belong to the same community.

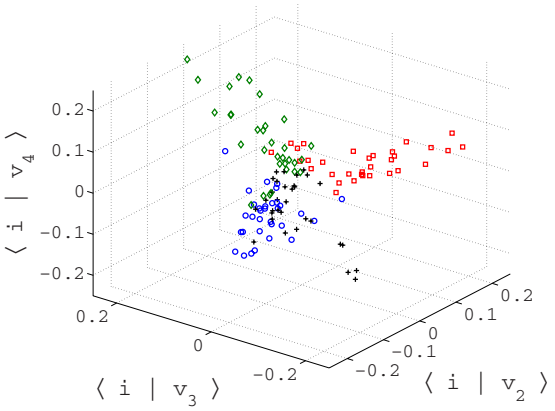

FIG. 7. (Color online) Plot of the eigenvector components of the second, third, and fourth eigenvectors of biased RWs with parameters $\beta_k=-2.5$ and $\beta_M=4.3$. Different markers represent four different predefined communities. This is an example of the same GN graph realization with $p_{in}=16/62$ and $p_{out}=1/12$ as the one on the previous figure. For this choice of parameters $\mu=1/2$. One can notice tetrahedral distribution of vertices in which vertices from the same community belong to the same branch of tetrahedron.

state. In numerical applications of such a theoretical framework we have observed a unimodal shape of the spectral gap vs the bias parameter, which is not an obvious feature of the studied processes. We have finally outlined a very promising application of topologically biased random walks to the fundamental problem of community finding. We described the basic ideas and proposed some criteria for the choice of parameters by considering the particular case of GN graphs. We are working further in direction of this application, but the number of possible strategies (different topological properties we can use for biasing) and types of networks are just too large to be presented in one paper. Furthermore, since in many dynamical systems such as the World Wide Web or biological networks feedback between function and form (topology) is evident, our framework may be a useful way to describe mathematically such an observed mechanism. In the case of biology, for instance, the shape of the metabolic networks can be triggered not only by the chemical properties of the compounds, but also by the possibility of the metabolites to interact. Biased RWs can be therefore the mechanism through which a network attains a particular form for a given function. By introducing such an approach we can now address the problem of community detection in the graph. This is the reason why here we have not introduced another precise method for community detection, but rather a possible framework to create different community finding methods with different *ad hoc* strategies. Indeed in real situations we expect different types of network to be efficiently explored by the use of different topological properties. This explains why we believe that TBRWs could play a role in community detection problems, and we hope to stimulate further developments, in the network scientific community, of this promising methodology.

ACKNOWLEDGMENTS

V.Z. would like to thank MSES of the Republic of Croatia through Project No. 098-0352828-2836 for partial support. The authors acknowledge support from EC FET Open Project "FOC" No. 255987.

[1] A. Broder, Comput. Netw. **33**, 309 (2000).
[2] V. Zlatić, M. Božičević, H. Štefančić, and M. Domazet, Phys. Rev. E **74**, 016115 (2006).
[3] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli, Phys. Rev. E **74**, 036116 (2006).
[4] M. Catanzaro, G. Caldarelli, and L. Pietronero, Phys. Rev. E **70**, 037101 (2004).
[5] J. B. Glattfelder and S. Battiston, Phys. Rev. E **80**, 036104 (2009).
[6] P. J. Cameron and C. Martíns, Combinatorics, Probab. Comput. **2**, 1 (1993).
[7] D. Aldous and J. Fill [http://stat-www.berkeley.edu/users/aldous/RWG/book.html].
[8] R. Pastor-Satorras and A. Vespignani, Phys. Rev. Lett. **86**, 3200 (2001).
[9] L. Page *et al.*, *The PAGERANK Citation Ranking: Bringing Order to the Web* (Stanford Digital Library Technologies Project, Stanford, CA, 1998).
[10] P. Blanchard and D. Volchenkov, *Mathematical Analysis of Urban Spatial Networks* (Springer, Berlin, 2009).
[11] S. Fortunato, Phys. Rep. **486**, 75 (2010).
[12] A. Capocci *et al.*, Physica A **352**, 669 (2005).
[13] L. Danon *et al.*, J. Stat. Mech.: Theory Exp. (2005) P09008.
[14] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[15] A. Arenas *et al.*, Phys. Rep. **469**, 93 (2008).
[16] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[17] M. Buchanan, *Nexus* (W.W. Norton & Co., New York, 2003).
[18] J. Gomez-Gardenes and V. Latora, Phys. Rev. E **78**, 065102(R) (2008).
[19] W. Feller, *An Introduction to Probability Theory and Its Applications* (John Wiley & Sons, New York, 1968).
[20] F. Chung, *Spectral Graph Theory*, CBMS Lecture Notes (AMS Publications, Providence, RI, 1992).
[21] J. D. Noh and H. Rieger, Phys. Rev. Lett. **92**, 118701 (2004).
[22] Z. Burda, J. Duda, J. M. Luck, and B. Waclaw, Phys. Rev. Lett. **102**, 160602 (2009).
[23] E. Pollak *et al.*, Biophys. J. **95**, 4258 (2008).
[24] D. A. Mazziotti *et al.*, J. Phys. Chem. **99**, 112 (1995).
[25] D. A. Mazziotti, Mol. Phys. **89**, 171 (1996).
[26] A. Lancichinetti and S. Fortunato, Phys. Rev. E **80**, 056117 (2009).
[27] A. Lancichinetti, S. Fortunato, and F. Radicchi, Phys. Rev. E **78**, 046110 (2008).
[28] A. Condon and R. M. Karp, Random Struct. Algorithms **18**, 116 (2001).
[29] L. Donetti and M. A. Muñoz, J. Stat. Mech.: Theory Exp. (2004) P10012.